文字の部分的な特徴量による主成分分析を用いた筆者の自動判定

田中 秀　　指導教員：何 立風

1 はじめに

遺言書などの筆者を識別する際に重要な証拠として筆跡は注目を浴びることが多いが、目視による識別には証明力に限界がある[1][2]。また、同じ人が同じ文字を書いた場合でも、全く同じ文字になるということはほぼなく、どこかに異なる部分が生じてしまう。これを変動性といい、筆者識別の難しさの原因の一つである[3]。

先行研究では筆圧に着目したものやクラスター分析を利用したものなど様々な方法があるが、本研究では筆者識別の精度向上のため、文字の部分的な特徴量に着目した判定方法を提案する[4][5]。

2 提案手法

本研究では 3 つの方法で筆者を判定する。

「判定方法 A」

判定方法 A は、文字を構成している特徴的な部分から様々な特徴量を抽出し、筆者が不明の文字と比較する方法である。手順としては、まずスキャナによって紙に書かれた手書き文字をデータ化し、それらの画像に対して二値化処理を施したのち、細線化することで文字の中心線を抜き出す[6]。次に「はね」や「はらい」などの文字を構成している特徴的な部分を抽出する。図 1 に細線化した画像と抽出した特徴部分の例を示す。なお、本研究で扱う特徴部分は縦線、横線、はね、はらい、点、かえり、まわり、輪、結び、かぎの 10 種類である。

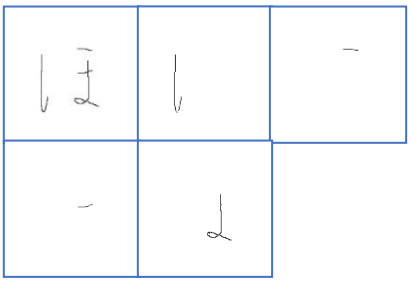

図 1 細線化画像と抽出した特徴部分

その後、すべての特徴部分から特徴量を抽出する。本研究で使用した特徴量は、高さ、幅、長さ、線の傾き、はねの角度、丸み、輪の大きさの 7 種類である。

筆者が判明している文字の特徴量を記録・更新する場合は、筆者ごとの各特徴部分の各特徴量に対してそれぞれ最大値と最小値を記録・更新する。筆者が不明の文字が、どの筆者が書いた文字であるかを判定する場合は、その文字から抽出した特徴量と、あらかじめ記録してある筆者の特徴量の最大値、最小値のデータを比較し、最大値より小さく最小値より大きければ 1、そうでない場合は 0 とする。1 文字ごとに全ての特徴量を比較し、その結果を合計していき、最終的に合計値を比較した特徴量の総数で除算し、100 倍して小数第 1 位で四捨五入したものを、その文字が比較した筆者と同一である確率(%)とする。

「判定方法 B」

判定方法 B は、判定方法 A によって抽出された特徴量から主成分分析によって新たな変数を作り出し、それらを比較する方法である。抽出された各筆者の特徴量ごとに平均値を求め、それらを標準化したものを$x_1, x_2, \cdots, x_n$とすると、各筆者について表 1 のようなデータが得られる。

表 1 各筆者の特徴量

	特徴量 *x1*	特徴量 *x2*	特徴量 *x3*	特徴量 *x4*	~	特徴量 *xn*
筆者A	*x*1A	*x*2A	*x*3A	*x*4A	~	*xn*A
筆者B	*x*1B	*x*2B	*x*3B	*x*4B	~	*xn*B
筆者C	*x*1C	*x*2C	*x*3C	*x*4C	~	*xn*C
筆者D	*x*1D	*x*2D	*x*3D	*x*4D	~	*xn*D

これらのデータに対して主成分分析を用いることで主成分 Z が求められる。その主成分 Z に各筆者の特徴量 x を代入することで筆者ごとに表 2 のような主成分得点が得られる。

表 2 各筆者の主成分得点

	主成分 Z1	主成分 Z2	主成分 Z3	主成分 Z4	~	主成分 Z*n*
筆者A	Z1A	Z2A	Z3A	Z4A	~	Z*n*A
筆者B	Z1B	Z2B	Z3B	Z4B	~	Z*n*B
筆者C	Z1C	Z2C	Z3C	Z4C	~	Z*n*C
筆者D	Z1D	Z2D	Z3D	Z4D	~	Z*n*D

筆者が不明の文字が与えられたとき、同じように主成分 Z の式に各特徴量を代入し主成分得点を求め、その値と最も近い値をもつ筆者をその文字の筆者とする。

「判定方法 C」

判定方法 C は、数値による解析とは異なり文字全体の大きさに対しての各特徴部分の位置に着目した方法である。

まず文字画像を黒い画素を内包する最小の正方形で切り取る。その後、図 2 のように画像を正規化し余白を加える。なお、今回は 50×50 の画像サイズに正規化した。

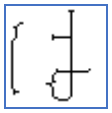

図 2 正規化後の画像

次に図 3 のように筆者別に黒い画素の数を画素の位置ごとにカウントしていく。そして、図 4 のような筆者不明の文字が与えられたとき、その画像の黒い画素と同じ位置にある筆者データの値を筆者別に合計していく。最終的に全ての筆者の得点を比較し、最も得点の高い筆者をその文字の筆者として判定する。

例：図 3,4 の場合、筆者 x の得点は
2+10+10+10+10+7+9+9+10+10+2+10+10+10+10+10+10=147
となる。

0	0	0	0	0	0	0	0
0	2	0	0	0	9	0	0
0	10	0	5	9	10	0	0
0	10	0	0	0	10	0	0
0	10	0	2	10	10	10	0
0	10	0	10	0	10	0	0
0	7	0	3	10	3	0	0
0	0	0	0	0	0	0	0

図 3 例：筆者 x の「は」のデータ
(10 文字分)

0	0	0	0	0	0	0	0
0	1	0	0	0	1	0	0
0	1	0	0	1	1	0	0
0	1	0	0	0	1	0	0
0	1	0	1	1	1	1	0
0	1	0	1	0	1	0	0
0	1	0	0	1	0	0	0
0	0	0	0	0	0	0	0

図 4 例：筆者が不明のデータ

3 実験

複数の人物によって書かれた文字に対して本研究の提案手法を実行し筆者の判定を行うことで、研究目的である文字の特徴部分に着目した筆者判定の精度の向上を図った。

被験者 8 人にボールペンで「て」、「に」、「は」の 3 種類の文字をそれぞれ 12 回ずつ書いてもらい、10 回分を筆者が判明している文字として特徴量を記録し、残りの 2 文字を筆者が不明の文字として、筆者ごとの特徴量のデータと比較し判定するために使用した。

4 考察

判定方法 A の実験結果は、被験者全体の正答率は約 70%だったが、被験者ごとに見るとすべての被験者において 50%を超えているが判定精度に大きな差が見られた。原因としては、学習データに対して使用するデータが少なかったこと、最大値と最小値の幅が大きい筆者ほど誤判定の可能性が高くなってしまうこと、1 文字でも字形の崩れたものがあるとその筆者の幅が大きくなってしまうことが挙げられる。

判定方法 B の実験結果は被験者全体の正答率が約 80%となり、判定方法 A に比べ判定精度の向上が見られた。その要因として、主成分分析を用いたことで判定方法 A での欠点であった最大値や最小値の幅による問題点が解消されたことが考えられる。また、判定方法 A とは異なり、与えられた学習データ全てを利用することができた点も要因の 1 つとして考えられる。

判定方法 C の実験結果は被験者全体の正答率が約 60%となり他の判定方法に比べ精度に欠ける結果となった。原因としては、正規化をしたことが挙げられる。判定方法 C では、各特徴部分の位置をある程度揃えるために正規化した画像を使用しなければならなかったが、複雑な文字であるほど画像を大きくする際に元の字形がくずれてしまう。正規化の精度を上げることで、元の字形を正確に保持したままより大きなサイズの画像を扱うことが可能になり、筆者の特徴をより表現できるようになるため判定精度が向上すると考えられる。

5 まとめ

全ての提案手法において一定の精度で筆者が正しく判定されるという結果が得られた。また、判定方法 C と比較して判定方法 A,B のほうが良い判定結果が得られたことから、筆者を判定するにあたって数値解析が有効であることが示された。

本研究では 3 種類の文字を使って判定を行ったが、実際に筆者判定を行う際にはより多くの種類の文字による判定を行うことが想定されるため、本研究の実験結果よりも精度の高い判定結果が得られることが考えられる。

本研究の提案手法による筆者判定の精度を上げるためには、細線化の精度を上げることで筆者の特徴を損なうことなく判定を行うことができるようにすること、数ある特徴量の中でどの特徴量が筆者を判定するうえで重要であるかを探し、特徴量ごとに重みをつけて判定すること、抽出する特徴量を増やし筆者の個人差をより多く見つけることなどが必要であると思われる。

参考文献

[1]吉村ミツ, 吉村功:”筆者認識研究の現段階と今後の動向”, 電子通信学会技術研究報告 PRMU, 1996(6), pp. 81-90, 1996-06-28
[2]利根, 相澤, 西川, 来代:”筆者識別指標の検討を支援するデータベースシステムの開発”, 法科学技術, 22(2), 145-151(2017)
[3]上田勝彦:”日本字筆跡の変動解析と筆跡個性に関する基礎的検討(第 2 報)”, 情報処理学会研究報告人文科学とコンピュータ(CH), 2008(8), pp. 27-34, 2008-01-25
[4]阿久津恵子, 笠村秀明, 古野晴彦:”文字の輪郭と曲線の特徴量に基づいた筆者の異同識別”, 画像の認識・理解シンポジウム(MIRU2011), 2011(7), pp. 525-528, 2011-07-20
[5]前川, 井俣, 大西:”ニューラルネットワークを用いた手書き文字の筆圧による個人識別”, 情報処理学会第 51 回(平成 7 年後期)
[6]高井信勝:”技術報告 2 値画像の細線化における連結数の役割の再検討”, 北海学園大学工学部研究報告, 2011(38), pp. 155-172, 2011-02-14